Claims

What is claimed is:

1. A method for representing statistics about a table including one or more rows, each row including a respective value, the method including:

creating zero or more histogram buckets, each histogram bucket including a width representing a respective range of values and a height representing a count of rows having values in the range of values; and

creating one or more high-bias buckets, each high-bias bucket representing one or more values that appear in a minimum percentage of rows.

2. The method of claim 1, where a total number of buckets is a fixed number equal to the sum of the number of histogram buckets and the number of high-bias buckets.

3. The method of claim 1, where creating the high-bias and histogram buckets includes:

(a) determining an average height of the histogram buckets;

(b) based on the average height of the histogram buckets, determining a reclassification threshold; and

(c) representing each value that exceeds the reclassification threshold in a high-bias bucket.

4. The method of claim 3, where the reclassification threshold is equal to the average height of the histogram buckets multiplied by (1+S), where S is a positive percentage represented as a decimal.

5. The method of claim 3, where (a), (b), and (c) are repeated until no value exceeds the reclassification threshold.

Express Mail No. EV339222752US
Date: December 31, 2003

6. The method of claim 1, where creating the high-bias and histogram buckets includes:

(a) determining an average height of the histogram buckets;

(b) based on the average height of the histogram buckets, determining a reclassification threshold; and

(c) for each value that exceeds the reclassification threshold:

(1) if all of the high-bias buckets are not full, representing the value in a high-bias bucket;

(2) else, if the number of high-bias buckets is less than a fixed number of high-bias buckets:

(i) creating a new high-bias bucket; and

(ii) representing the value in the new high-bias bucket.

7. The method of claim 6, where the reclassification threshold is equal to the average height of the histogram buckets multiplied by (1+S), where S is a positive percentage represented as a decimal.

8. The method of claim 6, where (a), (b), and (c) are repeated until:

(i) no value exceeds the reclassification threshold; or

(ii) a number of the high-bias buckets is equal to the fixed number of high-bias buckets and each of the high-bias buckets is full.

9. The method of claim 1, where a total number of buckets is equal to the sum of a number of histogram buckets and a number of high-bias buckets, where the total number of buckets is fixed, and where the method further includes:

(a) identifying one or more values that appear in at least the minimum percentage of rows and representing the identified values in the high-bias buckets;

(b) determining a remaining number of buckets equal to the total number of buckets less the number of high-bias buckets used; and

(c) if the number of remaining buckets is greater than a stop number of buckets:

(1) adjusting the minimum percentage of rows;

(2) identifying values that appear in the adjusted minimum percentage of rows; and

(3) representing values that appear in the adjusted minimum percentage of row in high-bias buckets.

Express Mail No. EV339222752US
Date: December 31, 2003

10. The method of claim 9, where (a) includes setting the minimum percentage of rows to $\frac{1}{FB}\%$, where F is equal to a number of high-bias values that each high-bias bucket can contain and B is equal to the total number of buckets.

11. The method of claim 9, where (c)(1) includes setting the adjusted minimum percentage to $\frac{V(FB-I)}{FB}\%$, where F is equal to a number of high-bias values that each high-bias bucket can contain, B is equal to the total number of buckets, V is equal to the minimum percentage of rows, and I is equal to a number of values represented in high-bias buckets.

12. The method of claim 9, further including:

(d) if the number of remaining buckets is less than or equal to the stop number of buckets:

representing values not represented in high-bias buckets in histogram buckets.

13. The method of claim 12, further including:

(e) repeating (b), (c), and (d) until the number of remaining buckets is less than or equal to the stop number of buckets.

14. The method of claim 1, where a total number of buckets is equal to the sum of a number of the histogram buckets and a number of the high-bias buckets, where the total number of buckets is fixed, where the number of high-bias buckets is fixed, and where the method includes:

populating the one or more high-bias buckets with the FH most frequently occurring values, where F is a number of values each high-bias bucket can store and H is the number of high-bias buckets; and

populating the one or more histogram buckets with all other values.

Express Mail No. EV339222752US
Date: December 31, 2003

15. A database system including:

a massively parallel processing system including:

one or more nodes;

a plurality of CPUs, each of the one or more nodes providing access to one or more CPUs;

a plurality of data storage facilities each of the one or more CPUs providing access to one or more data storage facilities;

P partitions, each partition residing on one or more data storage facilities;

a process for representing statistics, where the database system represents statistics about a table including one or more rows, each row including a respective value, the process including:

creating zero or more histogram buckets, each histogram bucket including a width representing a respective range of values and a height representing a count of rows having values in the range of values; and

creating one or more high-bias buckets, each high-bias bucket representing one or more values that appear in a minimum percentage of rows.

16. The database system of claim 15, where a total number of buckets is a fixed number equal to the sum of the number of histogram buckets and the number of high-bias buckets.

17. The database system of claim 15, where the process creating the high-bias and histogram buckets includes:

(a) determining an average height of the histogram buckets;

(b) based on the average height of the histogram buckets, determining a reclassification threshold; and

(c) representing each value that exceeds the reclassification threshold in a high-bias bucket.

18. The database system of claim 17, where the reclassification threshold is equal to the average height of the histogram buckets multiplied by (1+S), where S is a positive percentage represented as a decimal.

19. The database system of claim 17, where the process creating high-bias and histogram buckets includes repeating (a), (b), and (c) until no value exceeds the reclassification threshold.

Express Mail No. EV339222752US
Date: December 31, 2003

20. The database system of claim 15, where the process creating high-bias and histogram buckets includes:

(a) determining an average height of the histogram buckets;

(b) based on the average height of the histogram buckets, determining a reclassification threshold; and

(c) for each value that exceeds the reclassification threshold:

(1) if all of the high-bias buckets are not full, representing the value in a high-bias bucket;

(2) else, if the number of high-bias buckets is less than a fixed number of high-bias buckets:

(i) creating a new high-bias bucket; and

(ii) representing the value in the new high-bias bucket.

21. The database system of claim 20, where the reclassification threshold is equal to the average height of the histogram buckets multiplied by (1+S), where S is a positive percentage represented as a decimal.

22. The database system of claim 20, where the process creating high-bias and histogram buckets repeats (a), (b), and (c) until:

(i) no value exceeds the reclassification threshold; or

(ii) a number of the high-bias buckets is equal to the fixed number of high-bias buckets and each of the high-bias buckets is full.

23. The database system of claim 15, where a total number of buckets is equal to the sum of a number of histogram buckets and a number of high-bias buckets, where the total number of buckets is fixed, and where the process creating the high-bias and histogram buckets further includes:

(a) identifying one or more values that appear in at least the minimum percentage of rows and representing the identified values in the high-bias buckets;

(b) determining a remaining number of buckets equal to the total number of buckets less the number of high-bias buckets used; and

(c) if the number of remaining buckets is greater than a stop number of buckets:

(1) adjusting the minimum percentage of rows;

(2) identifying values that appear in the adjusted minimum percentage of rows; and

(3) representing values that appear in the adjusted minimum percentage of row in high-bias buckets.

Express Mail No. EV339222752US
Date: December 31, 2003

24. The database system of claim 23, where (a) includes setting the minimum percentage of rows to $\frac{1}{FB}\%$, where F is equal to a number of high-bias values that each high-bias bucket can contain and B is equal to the total number of buckets.

25. The database system of claim 23, where (c)(1) includes setting the adjusted minimum percentage to $\frac{V(FB-I)}{FB}\%$, where F is equal to a number of high-bias values that each high-bias bucket can contain, B is equal to the total number of buckets, V is equal to the minimum percentage of rows, and I is equal to a number of values represented in high-bias buckets.

26. The database system of claim 23, where the process creating the high-bias and histogram buckets further includes:

(d) if the number of remaining buckets is less than or equal to the stop number of buckets:

representing values not represented in high-bias buckets in histogram buckets.

27. The database system of claim 26, where the process creating the high-bias and histogram buckets further includes:

(e) repeating (b), (c), and (d) until the number of remaining buckets is less than or equal to the stop number of buckets.

28. The database system of claim 15, where a total number of buckets is equal to the sum of a number of the histogram buckets and a number of the high-bias buckets, where the total number of buckets is fixed, where the number of high-bias buckets is fixed, and where , where the process creating the high-bias and histogram buckets further includes:

populating the one or more high-bias buckets with the FH most frequently occurring values, where F is a number of values each high-bias bucket can store and H is the number of high-bias buckets; and

populating the one or more histogram buckets with all other values.

29. A computer program, stored on a tangible storage medium, for use in representing statistics in a database running in a partitioned parallel environment including P partitions, each partition residing on one or more parallel processing systems, the database including a first table including one or more rows stored in one or more of the P partitions, the program including executable instructions that cause a computer to:

represent statistics about a table including one or more rows, each row including one or more values, the program further causing the computer to:

create zero or more histogram buckets, each histogram bucket including a width representing a respective range of values and a height representing a count of rows having values in the range of values; and

create one or more high-bias buckets, each high-bias bucket representing one or more values that appear in a minimum percentage of rows.

30. The computer program of claim 29, where a total number of buckets is a fixed number equal to the sum of the number of histogram buckets and the number of high-bias buckets.

31. The computer program of claim 29, including executable instructions that cause the computer to:

(a) determine an average height of the histogram buckets;

(b) based on the average height of the histogram buckets, determine a reclassification threshold; and

(c) represent each value that exceeds the reclassification threshold in a high-bias bucket.

32. The computer program of claim 31, where the reclassification threshold is equal to the average height of the histogram buckets multiplied by (1+S), where S is a positive percentage represented as a decimal.

33. The computer program of claim 31, including executable instructions that cause the computer to repeat (a), (b), and (c) until no value exceeds the reclassification threshold.

34. The computer program of claim 29, including executable instructions that cause the computer to:

(a) determine an average height of the histogram buckets;

(b) based on the average height of the histogram buckets, determine a reclassification threshold; and

(c) for each value that exceeds the reclassification threshold:

(1) if all of the high-bias buckets are not full, represent the value in a high-bias bucket;

(2) else, if the number of high-bias buckets is less than a fixed number of high-bias buckets:

(i) create a new high-bias bucket; and

(ii) represent the value in the new high-bias bucket.

35. The computer program of claim 34, where the reclassification threshold is equal to the average height of the histogram buckets multiplied by (1+S), where S is a positive percentage represented as a decimal.

36. The computer program of claim 34, including executable instructions that cause the computer to repeat (a), (b), and (c) until:

(i) no value exceeds the reclassification threshold; or

(ii) a number of the high-bias buckets is equal to the fixed number of high-bias buckets and each of the high-bias buckets is full.

37. The computer program of claim 29, where a total number of buckets is equal to the sum of a number of histogram buckets and a number of high-bias buckets, where the total number of buckets is fixed, and where the computer program includes executable instructions that cause the computer to:

(a) identify one or more values that appear in at least the minimum percentage of rows and representing the identified values in the high-bias buckets;

(b) determine a remaining number of buckets equal to the total number of buckets less the number of high-bias buckets used; and

(c) if the number of remaining buckets is greater than a stop number of buckets:

(1) adjust the minimum percentage of rows;

(2) identify values that appear in the adjusted minimum percentage of rows; and

(3) represent values that appear in the adjusted minimum percentage of row in high-bias buckets.

Express Mail No. EV339222752US
Date: December 31, 2003

38. The computer program of claim 37, where (a) includes further executable instructions that cause the computer to set the minimum percentage of rows to $\frac{1}{FB}\%$, where F is equal to a number of high-bias values that each high-bias bucket can contain and B is equal to the total number of buckets.

39. The computer program of claim 37, where (c)(1) includes further executable instructions that cause the computer to set the adjusted minimum percentage to $\frac{V(FB-I)}{FB}\%$, where F is equal to a number of high-bias values that each high-bias bucket can contain, B is equal to the total number of buckets, V is equal to the minimum percentage of rows, and I is equal to a number of values represented in high-bias buckets.

40. The computer program of claim 37, further including executable instructions that cause the computer to:

(d) if the number of remaining buckets is less than or equal to the stop number of buckets:

represent values not represented in high-bias buckets in histogram buckets.

41. The computer program of claim 40, further including executable instructions that cause the computer to:

(e) repeat (b), (c), and (d) until the number of remaining buckets is less than or equal to the stop number of buckets.

42. The computer program of claim 29, where a total number of buckets is equal to the sum of a number of the histogram buckets and a number of the high-bias buckets, where the total number of buckets is fixed, where the number of high-bias buckets is fixed, and where the computer program includes executable instructions that cause the computer to:

populate the one or more high-bias buckets with the FH most frequently occurring values, where F is a number of values each high-bias bucket can store and H is the number of high-bias buckets; and

populate the one or more histogram buckets with all other values.